Edge-Deployable Dual-Branch Network with Cross-Attention for Multi-Source Gait Recognition

Ruixiang Hu[1], Chenggang Lu[2], Zhengqing He[3], Yan Wang[4], Hengyi Li[5], Yuguo Chen[6], Hongnian Yu[7]
{ruixianghu@zut.edu.cn[1], luchenggang@zut.edu.cn[2], zhengqinghe@zut.edu.cn[3]}

Zhongyuan University of Technology, No.1 Huaihe Road, Zhengzhou, Henan, 451191, China[1,2,3]

Abstract. Human gait serves as an important indicator for neuromuscular, musculoskeletal, and cognitive health, and its accurate recognition is critical in health monitoring, rehabilitation assessment, and identity verification. Single-sensor approaches, however, often fail to capture the complex spatiotemporal dynamics of gait. To address this limitation, we propose DBTCNet, a cross-branch attention network that fuses plantar pressure and inertial measurement unit (IMU) data for robust gait recognition. Each branch independently extracts spatial and temporal features from its modality through convolutional and temporal attention modules, while a cross-branch attention mechanism explicitly models inter-modal dependencies, enhancing complementary and discriminative feature representations. Evaluations on a self-collected dataset covering 11 gait patterns show that DBTCNet achieves 97.81% accuracy, outperforming the best single-modality approach by 0.96% in F1-score. The model was further deployed on a Raspberry Pi 4B, achieving an average inference latency of 32.14 ms with 98% accuracy, meeting real-time requirements for edge-based abnormal gait classification.

Keywords: gait recognition; multi-source data fusion; dual-branch network; edge deployment; wearable sensors

1 Introduction

Gait is a periodic, rhythmic pattern of human locomotion in three-dimensional space that reflects an individual's motor control strategy and overall neuromuscular, musculoskeletal, and cognitive health status [1, 2]. Clinical studies indicate that disorders such as Parkinson's disease, stroke, and musculoskeletal injuries can cause substantial alterations in gait spatiotemporal parameters, kinetic characteristics, and movement coordination [3]. Accurate and timely gait recognition is therefore critical for early screening, disease progression monitoring, rehabilitation assessment, and the formulation of personalized intervention strategies.

The rapid advancement of wearable sensing technologies, such as inertial measurement units (IMU) and plantar pressure insoles, has enabled high-resolution, multimodal gait data acquisition in

IIKI 2025, December 18-21, Chengdu, People's Republic of China

DOI 10.4108/eai.18-12-2025.2365291

unconstrained environments [4, 5]. These modalities are inherently complementary: IMU captures limb kinematics and dynamics [6], while plantar pressure sensors provide ground reaction forces and plantar load distribution [7]. However, gait recognition approaches relying on a single modality often exhibit reduced robustness under complex walking conditions or in abnormal gait scenarios [8, 9].

For this reason, many researchers have explored multi-source sensor methods. Previous studies have demonstrated that the complementary information provided by multimodal data can significantly enhance system representational capacity [10]. Meanwhile, several wearable platforms have been developed that integrate plantar pressure insoles, inertial measurement units, and additional sensing modalities, thereby validating the advantages of multi-source sensors in complex motion scenarios [11, 12]. Building on these advances, multimodal data fusion has emerged as a critical strategy for further improving system performance, achieving significant gains in applications such as disease monitoring and gait assessment [13, 14].

In parallel, attention mechanisms have been increasingly adopted to capture multi-dimensional features within unimodal data [15]. For example, [16] proposed a cross-attention module for plantar pressure signals to jointly model temporal, channel, and spatial dependencies. [17] introduced a three-branch attention framework to extract temporal, spatial, and spatiotemporal representations from multi-sensor inputs. Although such multi-branch parallel architectures can substantially enhance feature extraction capability [18], they inevitably increase model complexity. Moreover, these methods primarily focus on intra-modality feature modeling and fail to explicitly capture dependencies across different modalities.

To address the limitations, we propose DBTCNet, a dual-branch temporal and cross-attention network designed for robust gait recognition from multimodal wearable data. In DBTCNet, plantar pressure and IMU data are processed in parallel branches to preserve modality-specific characteristics, while temporal attention modules enhance intra-modality temporal modeling. A cross-attention mechanism is introduced to explicitly capture inter-modal dependencies, enabling each branch to selectively integrate complementary information from the other modality. Furthermore, DBTCNet employs lightweight convolutional module and efficient attention structures, making it suitable for real-time deployment on edge devices. We validate DBTCNet on a self-collected dataset covering 11 gait patterns, and deploy it on a Raspberry Pi 4B [19, 20] to assess its feasibility in wearable and clinical scenarios.

The main contributions of this work are as follows:

1) We design a dual-branch architecture to separately extract modality-specific features from plantar pressure and IMU data, integrating temporal attention and cross-branch attention to explicitly model intra- and inter-modal dependencies.

2) We implement DBTCNet on a Raspberry Pi 4B, achieving an average inference latency of 32.14 ms with 98% accuracy, meeting real-time and low-power requirements for wearable and rehabilitation monitoring scenarios.

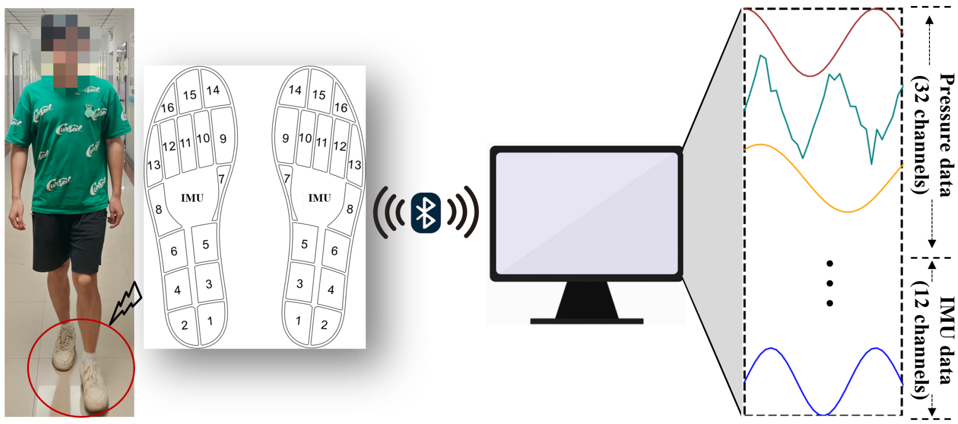

Fig. 1. Schematic diagram of the data collection.

2 DATASET

In this study, gait data were collected from 12 participants (7 males and 5 females, aged 22-26) using Opengo smart insoles. Each insole was equipped with 16 pressure sensors (covering forefoot, midfoot, and heel regions) and a 6-axis IMU (triaxial accelerometer + triaxial gyroscope), sampling at 100 Hz. The dataset comprised 11 gait types: four normal gaits (Slow Walking, Normal Walking, Fast Walking, Running) and seven abnormal gaits (In-toeing, Out-toeing, Right-sided Antalgic, Left-sided Antalgic, Magnetic, Steppage, and Gluteus Medius Gait).

Fig.1 illustrates the schematic diagram of the data collection. Prior to data collection, participants underwent pre-trial gait training to ensure correct execution of each gait type. For each trial, approximately two minutes of gait data were recorded per gait type, with a minimum two-minute rest interval provided between trials to mitigate fatigue effects. Data acquisition was conducted under the supervision of two researchers to ensure accuracy and annotation consistency. In total, 1,584,000 gait samples were collected.

3 THE PROPOSED DBTCNet

As illustrated in Fig.2, the proposed dual-branch network (DBTCNet) is designed to exploit the complementary nature of plantar pressure and IMU signals while preserving their modality-specific characteristics. The architecture consists of four main components: (1) a convolutional module that captures local spatial features within each modality; (2) an intra-branch attention module that models long-range temporal dependencies; (3) a cross-branch attention module that enables selective exchange of complementary cues across modalities; and (4) a feature concatenation and classification module that integrates enriched features for final gait category prediction. In the attention modules, Q, K, and V denote the query, key, and value matrices derived from the convolutional features of the same branch (intra-branch) or from the other branch (cross-branch, denoted Q_1, K_1, V_1

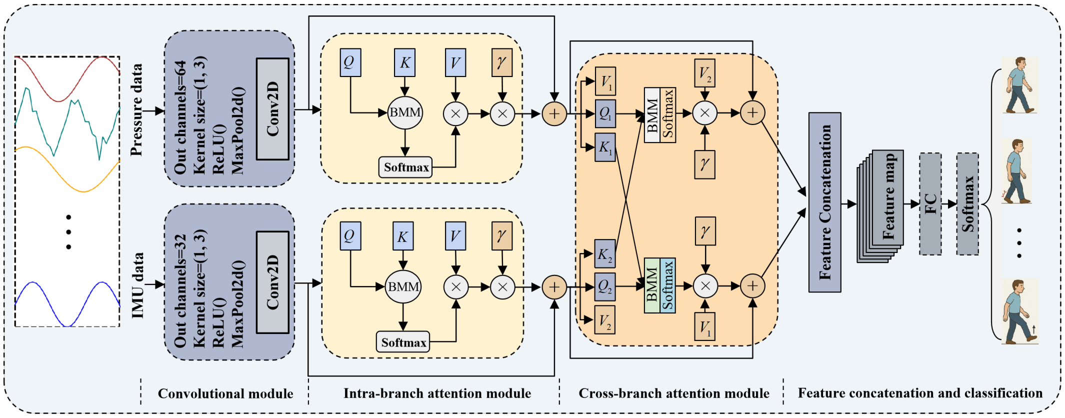

Fig. 2. Architecture of the proposed dual-branch network with temporal and cross-branch attention modules. BMM: Batch Matrix Multiplication.

and Q_2, K_2, V_2). The batch matrix multiplication (BMM) operation computes temporal similarity between Q and K, while a learnable scalar γ adjusts the fusion ratio between the attention output and the original features, retaining more original features when $\gamma \to 0$ and increasing attention influence when γ is larger.

3.1 Intra-branch Attention Module

While convolutional neural networks excel at extracting local spatial patterns, they have limited capability in capturing long-range temporal dependencies, an essential factor in gait sequences where discriminative cues may occur at different phases of the gait cycle. To address this, each branch in DBTCNet incorporates an intra-branch attention module to enhance temporal modeling within a single modality, ensuring that key temporal patterns are preserved before cross-modal fusion. Given the convolutional feature maps $X \in R^{B\times C\times 1\times T}$ (batch size B, channel dimension C , temporal length T), three parallel 1×1 convolutions generate the query Q, key K, and value V matrices:

$$Q = f_q(X), \quad K = f_k(X), \quad V = f_v(X) \tag{1}$$

After convolution and removing the height dimension, we obtain $Q, K \in R^{B\times C_q\times T}$ and $V \in R^{B\times C\times T}$. The temporal attention weights are then computed using the scaled dot-product attention [21]:

$$Attn = Softmax\left(\frac{Q^{\mathrm{T}}K}{\sqrt{C_q}}\right), \quad \mathrm{Attn} \in R^{B\times T\times T} \tag{2}$$

These weights quantify the similarity between different time steps, allowing the module to focus on temporally relevant regions of the gait cycle. The weighted sum of V produces the attended feature map:

$$O = V \cdot \mathrm{Attn}^{\mathrm{T}}, \quad O \in R^{B\times C\times T} \tag{3}$$

Finally, a learnable scalar parameter γ is introduced, which adaptively adjusts its value during training according to the optimization objective, thereby dynamically modulating the contribution of the attention outputs:

$$Y = \gamma O + X \tag{4}$$

3.2 Cross-branch Attention Module

Following the intra-branch attention, DBTCNet employs a cross-branch attention module to explicitly model dependencies between the two modalities and enhance the complementarity of fused features. As shown in Fig.2, this module takes the intermediate outputs from the plantar pressure branch and the IMU branch, and performs bidirectional information exchange.

Specifically, the intermediate outputs of the pressure branch and the IMU branches are used to construct Q_1, K_1, V_1 and Q_2, K_2, V_2, respectively. In the pressure branch, Q_1 is multiplied with K_2 through batch matrix multiplication, followed by Softmax normalization to obtain the attention weights. These weights are then multiplied with V_2 from the IMU branch to generate the weighted feature map. Similarly, in the IMU branch, Q_2 and K_1 undergo the same operations, and the result is multiplied with V_1 to produce the branch output. The outputs of both branches are subsequently scaled by a learnable parameter γ and added to the input of the cross-branch attention module to yield the final output.

Structurally, the attention computation follows the same mechanism as the intra-branch attention module, with the key difference being that Q, K, and V are derived from different branches rather than the same one. This shift enables the module to move beyond temporal dependency modeling within a modality toward capturing cross-modal correlations. By integrating complementary cues in this targeted manner, the cross-branch attention module facilitates more fine-grained and discriminative multi-source fusion with simple concatenation or averaging.

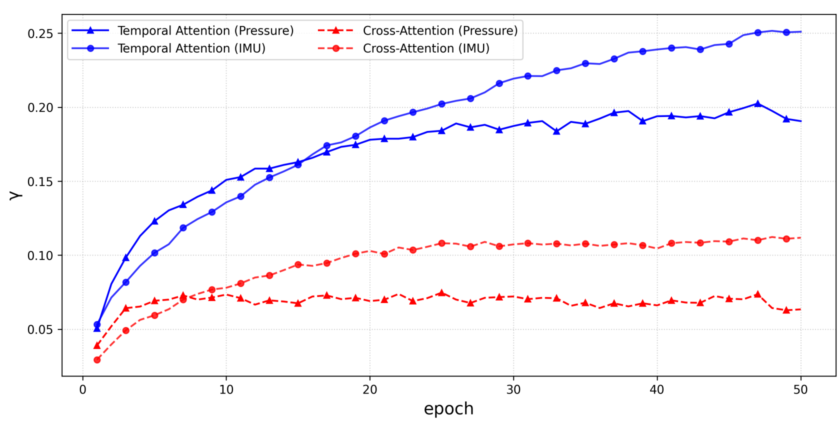

Fig. 3. Variation of attention weights (γ) over training epochs. Temporal attention and cross-branch attention respectively illustrate the evolution of γ within different attention modules, while pressure and IMU correspond to the variation of γ across distinct data branches.

4 EXPERIMENTAL RESULTS

For the experiments, the raw data consisted of 32-channel plantar pressure signals and 12-channel IMU signals. All data were normalized using the min-max method to eliminate dimensional differences. A sliding segmentation approach was then applied, with both the window length and step size set to 128, resulting in 12,375 windowed samples without overlapping. Finally, a ten-fold cross-validation strategy was employed to train and evaluate the model.

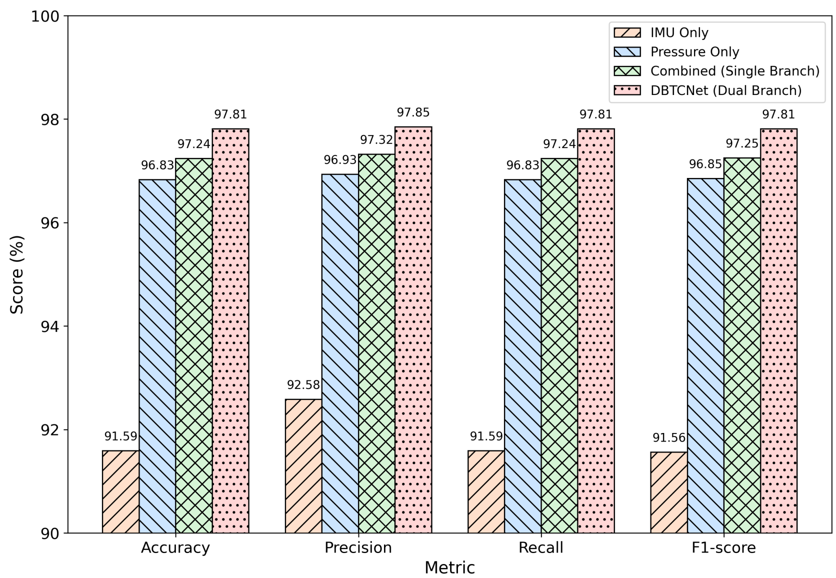

Fig. 4. Performance comparison across different input data modalities.

4.1 Evolution of Attention Weights γ Across Training

Fig.3 illustrates the variation of the attention weight coefficients γ across training epochs for different attention modules in DBTCNet and data modalities. In all cases, γ exhibits a steadily indicating progressive strengthening of attention weights during training. Overall, the values of γ in the temporal attention are consistently higher than those in the cross-branch attention module, suggesting that temporal dependencies contribute more substantially to capturing discriminative gait features. Furthermore, within the same attention mechanism, IMU signals yield higher γ values than pressure signals, implying that IMU data provide stronger and more stable cues for gait representation.

Table 1: Performance comparison of different configurations

Configuration	Evaluation Metrics				Model Complexity (M)	
	Accuracy	Precision	Recall	F1-score	FLOPs	Params
Conv2D only	95.80%	95.92%	95.80%	95.80%	8.04	3.11
Conv2D + IA	97.32%	97.39%	97.32%	97.32%	9.38	3.12
Conv2D + CA	97.35%	97.42%	97.35%	97.35%	8.70	3.12
IA + CA	95.69%	95.82%	95.69%	95.70%	7.06	2.89
Conv2D + IA + CA	**97.81%**	**97.85%**	**97.81%**	**97.81%**	10.04	3.13

Note: Conv2D = 2D Convolution Module, IA = Intra-branch Attention Module, CA = Cross-branch Attention Module.

4.2 Ablation Study on Input Modalities

As shown in Fig.4, IMU Only and Pressure Only use a single modality, while Combined (Single Branch) and DBTCNet (Dual Branch) fuse both modalities. Among them, IMU Only, Pressure Only, and Combined (Single Branch) adopt the single-branch architecture in Fig.5, with the dual-branch structure and cross-attention module removed for direct comparison. Results indicate that plantar pressure data achieve higher recognition performance (F1-score = 96.85%) than IMU data alone (F1-score = 91.56%), suggesting greater discriminative power in gait recognition. Combining both modalities further enhances performance, with the proposed DBTCNet achieving the highest F1-score (97.81%), thereby validating the effectiveness of the dual-branch design.

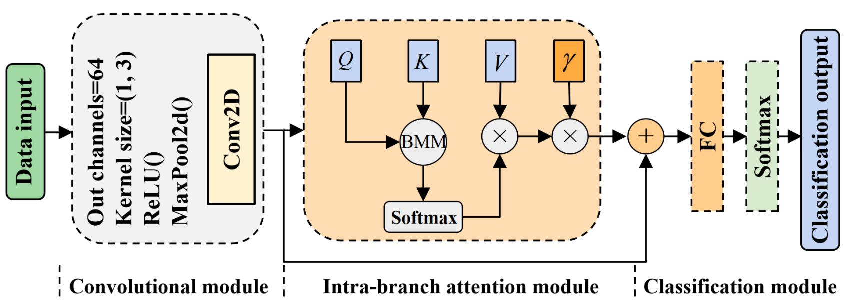

Fig. 5. Single-branch network model.

4.3 Comparison of Different Data Input Methods

To assess the dual-branch model's adaptability under different input configurations, two strategies were compared. In the first, plantar pressure and IMU data were fed into separate branches for feature extraction and fusion. In the second, data from the left and right foot were assigned to separate branches. The results (Fig.6) show that the first strategy (solid line) achieves superior performance in most gait categories, notably in Normal Walking, Steppage Gait, and Gluteus Medius Gait. However, the second strategy (dashed line) matches or exceeds performance in antalgic gaits affecting both legs, such as Right- and Left-sided Antalgic Gait, where pronounced left-right asymmetry enhances cross-branch complementarity. These findings confirm that allocating complementary data to distinct branches can improve the discrimination of diverse gait patterns.

4.4 Module Contribution Analysis

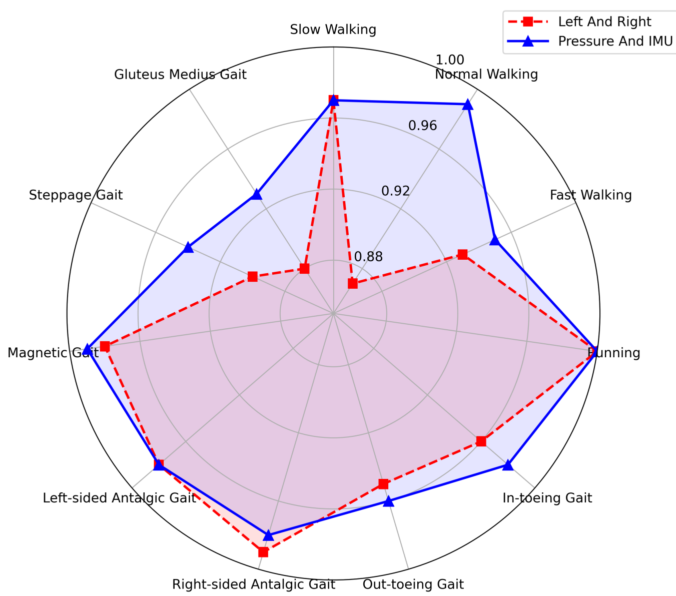

Fig. 6. Performance comparison of two input partitioning strategies: separating plantar pressure and IMU data (solid line) versus separating left and right foot data (dashed line) across multiple gait categories.

Table I summarizes the results, highlighting the contribution of each module to gait recognition. Using only the 2D Convolution module yields 95.80% accuracy, indicating limited capacity to capture temporal dependencies and cross-modal correlations. Adding Temporal Attention or Cross-Attention improves accuracy to 97.32% and 97.35%, respectively, an increase of 1.5% over the 2D Convolution, demonstrating the benefits of temporal modeling and cross-modal interaction. Removing convolution reduces accuracy to 95.69%, confirming its importance for local feature extraction. Combining all three modules achieves the best performance (Accuracy 97.81%, Precision 97.85%, Recall 97.81%, F1-score 97.81%) but with slightly higher complexity (FLOPs 10.04M, Params 3.13M).

As shown in Fig.7, the feature heatmaps of DBTCNet at different stages reveal the progressive

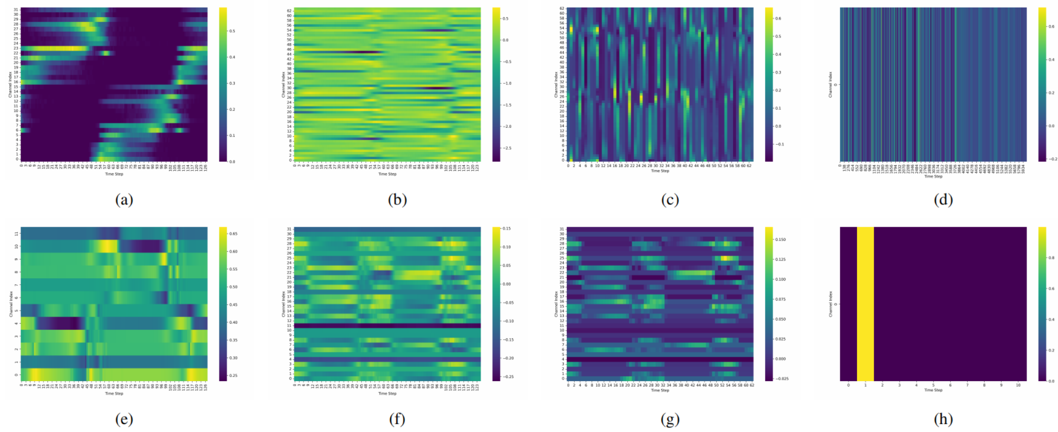

Fig. 7. Feature heatmaps of DBTCNet at different stages for normal walking: (a, e) raw inputs; (b, f) post-convolution features; (c, g) post-temporal attention features; (d) flattened features; (h) Softmax output probabilities.

refinement of feature representations. In the heatmaps, brighter colors denote larger feature values, indicating that specific channels or time steps receive greater attention, whereas darker colors denote weaker activations. The raw inputs initially present irregular temporal fluctuations; after convolution, more structured local patterns emerge; during the temporal attention stage, the model further emphasizes discriminative temporal segments. Following feature flattening, the representations are reorganized into compact vectors, enabling the classifier to emphasize salient components. Finally, at the Softmax output stage, the activations become highly sparse, with the network's response concentrated on the correct class.

4.5 Generalization Analysis via LOSO Cross-Validation

To assess cross-subject generalization, DBTCNet was evaluated using leave-one-subject-out (LOSO) cross-validation. As shown in Table II, performance varied notably across subjects: Subject 3 achieved the highest F1-score (80.36%), whereas Subject 10 had the lowest (53.59%), yielding a range of 26.77%. Most subjects' accuracies fell within the range of 65%-75%, suggesting that the model achieves a moderate level of cross-subject generalization. Nevertheless, notable inter-individual differences remain: even for the same gait type, subjects exhibit substantial variations in gait patterns, which in turn cause shifts in feature distributions. These differences pose inherent challenges to achieving robust cross-subject generalization, highlighting the need for advanced strategies such as subject-invariant feature learning or domain adaptation to further improve performance.

Table 2: LOSO results of DBTCNet across test subjects

Test Subject	Evaluation Metrics (%)			
	Acc.	Pr.	Re.	F1-score
Subject 1	71.91	78.17	71.91	71.21
Subject 2	63.68	77.55	63.68	58.09
Subject 3	81.37	87.89	81.37	80.36
Subject 4	59.27	68.89	59.27	54.77
Subject 5	66.85	69.43	66.85	64.78
Subject 6	72.88	78.38	72.88	70.71
Subject 7	67.56	74.20	67.56	65.13
Subject 8	72.04	83.66	72.04	67.42
Subject 9	58.68	66.89	58.68	55.74
Subject 10	56.78	60.61	56.78	53.59
Subject 11	72.11	77.33	72.11	72.57
Subject 12	70.81	67.55	70.81	67.14

Note: Acc. = Accuracy, Pr. = Precision, Re. = Recall.

4.6 Raspberry Pi 4B Deployment

To evaluate the proposed DBTCNet's deployability and real-time performance, the trained model was implemented on a Raspberry Pi 4B (8 GB RAM, Raspberry Pi OS 64-bit, Python 3.7.3, PyTorch 1.9) for edge inference. One hundred data windows were randomly selected from all samples for testing. Fig.8 shows the human-computer interaction interface, featuring a modular de-

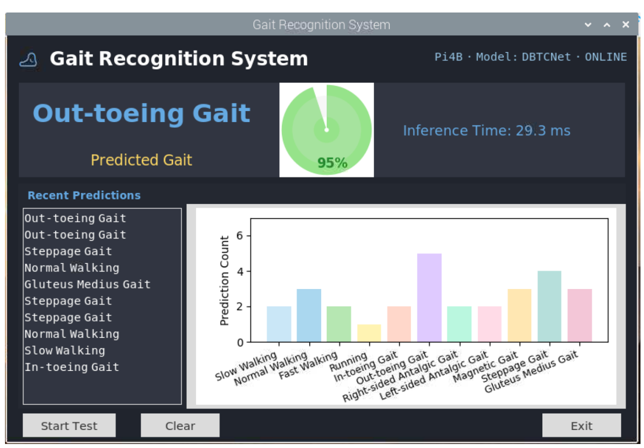

Fig. 8. User interface of the gait recognition system running on Raspberry Pi 4B.

sign: the upper panel displays the predicted gait type, confidence score, and inference time; the lower panel provides recent predictions (left) and a frequency distribution of detected gait categories (right). Fig.9 presents per-sample inference results, with blue circles indicating correct predictions and red triangles indicating incorrect predictions, and a green dashed line denoting the average inference time. The system achieved 98% accuracy with an average inference time of 32.14 ms, while maintaining stable memory usage at 222.27 MB. These results demonstrate the efficiency of the proposed framework in real-time scenarios. It should be emphasized, however, that this evaluation was conducted on a limited sample for inference benchmarking, and thus complements rather than replaces the comprehensive offline accuracy evaluation.

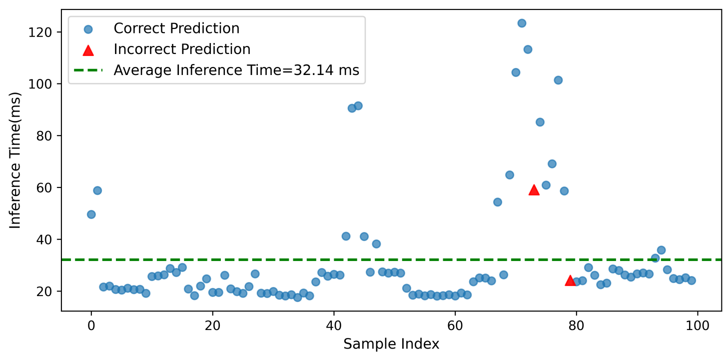

Fig. 9. Inference results for each sample on the Raspberry Pi 4B.

5 CONCLUSION

This study proposed a dual-branch network for multi-source gait recognition, effectively integrating multi-channel plantar pressure data and inertial measurement unit signals. The architecture employs parallel branches to extract temporal dependencies from pressure and IMU data respectively, and explicitly models inter-modal dependencies via a cross-branch attention mechanism, thereby achieving adaptive fusion of complementary cross-modal information. On the self-collected dataset comprising 12 subjects and 11 distinct gait types, the model achieved an accuracy of 97.81%, surpassing single-modality baselines and confirming the advantages of spatiotemporal feature fusion. Cross-subject generalization was evaluated via leave-one-subject-out validation. Although deployment on the resource-constrained Raspberry Pi 4B edge device confirmed the real-time inference feasibility of the system, LOSO experimental results revealed substantial discrepancies in data distribution across individuals, indicating that the model tends to automatically learn subject-specific idiosyncrasies during the training phase, thereby resulting in insufficient generalization capability when confronted with novel individuals. Therefore, future work will focus on enhancing cross-subject generalization and computational efficiency. Domain adaptation techniques will be introduced to align feature distributions across individuals, physical prior features will be incorporated to guide the model toward cross-subject stable discriminative patterns, and structured pruning alongside quantization methods will be applied to compress the model footprint, thereby reducing memory overhead and supporting sustained, efficient edge deployment.

Acknowledgments. This research was partially funded by the Key International Science and Technology Cooperation Project of Henan Province (Grant No. 251111520400), the ZUT Graduate Research Innovation Program (Grant No. YKY20252K06), and the Henan Provincial Science and Technology Research and Development Joint Fund Project (Grant No. 252103810251). Additional support was provided by the Henan Province Key Research and Development Projects (Grant Nos. 241111312000, 25210221110, 242102211066, and 252102320281).

References

[1] Liu X, Li W, Liu Z, Du F, Zou Q. A dual-branch model for diagnosis of Parkinson's disease based on the independent and joint features of the left and right gait. Applied Intelligence. 2021;51(10):7221-32.

[2] De Groote F, Falisse A. Perspective on musculoskeletal modelling and predictive simulations of human movement to assess the neuromechanics of gait. Proceedings of the Royal Society B. 2021;288(1946):20202432.

[3] Shin SY, Kim Y, Jayaraman A, Park HS. Relationship between gait quality measures and modular neuromuscular control parameters in chronic post-stroke individuals. Journal of neuroengineering and rehabilitation. 2021;18(1):58.

[4] Santos VM, Gomes BB, Neto MA, Amaro AM. A systematic review of insole sensor technology: recent studies and future directions. Applied Sciences. 2024;14(14):6085.

[5] Wang H, Wang X, Lu C, Yuan M, Wang Y, Yu H, et al. Enhancing Human Activity Recognition in Wrist-Worn Sensor Data Through Compensation Strategies for Sensor Displacement. IEEE Access. 2024;12:95058-70.

[6] Malleson C, Collomosse J, Hilton A. Real-time multi-person motion capture from multi-view video and IMUs. International Journal of Computer Vision. 2020;128(6):1594-611.

[7] Oubre B, Lane S, Holmes S, Boyer K, Lee SI. Estimating ground reaction force and center of pressure using low-cost wearable devices. IEEE transactions on biomedical engineering. 2021;69(4):1461-8.

[8] Yaprak B, Gedikli E. Different gait combinations based on multi-modal deep CNN architectures. Multimedia Tools and Applications. 2024;83(35):83403-25.

[9] Liu J, Xu X, Qiu Y, Wang C. Dual-stream interactive mechanism with multi-modal hierarchical aggregation transformer for gait recognition. Scientific Reports. 2025;15(1):26079.

[10] Lin K, Li Y, Sun J, Zhou D, Zhang Q. Multi-sensor fusion for body sensor network in medical human–robot interaction scenario. Information Fusion. 2020;57:15-26.

[11] Salis F, Bertuletti S, Bonci T, Caruso M, Scott K, Alcock L, et al. A multi-sensor wearable system for the assessment of diseased gait in real-world conditions. Frontiers in Bioengineering and Biotechnology. 2023;11:1143248.

[12] Liu J, Tan X, Jia X, Li T, Li W. A gait phase recognition method for obstacle crossing based on multi-sensor fusion. Sensors and Actuators A: Physical. 2024;376:115645.

[13] Tang L, Hu Q, Wang X, Liu L, Zheng H, Yu W, et al. A multimodal fusion network based on a cross-attention mechanism for the classification of Parkinsonian tremor and essential tremor. Scientific Reports. 2024;14(1):28050.

[14] Liu X, Zhang X, Li J, Pan W, Sun Y, Lin S, et al. Automated UPDRS Gait Scoring Using Wearable Sensor Fusion and Deep Learning. Bioengineering. 2025;12(7):686.

[15] Wang Y, Wang X, Yang H, Geng Y, Yu H, Zheng G, et al. MhaGNN: A novel framework for wearable sensor-based human activity recognition combining multi-head attention and graph neural networks. IEEE Transactions on Instrumentation and Measurement. 2023;72:1-14.

[16] Yuan M, Wang Y, Zhou X, Gui M, Wang A, Wang C, et al. Foot Pressure-Based Abnormal Gait Recognition With Multi-Scale Cross-Attention Fusion. IEEE Transactions on Neural Systems and Rehabilitation Engineering. 2025.

[17] Yang H, Wang Y, Hu R, Geng Y, Wang A, Zhou X, et al. Enhancing human activity recognition with TB-ConvAtt: A multi-dimensional attention framework. Biomedical Signal Processing and Control. 2025;110:108314.

[18] Liu J, Wang Q, Fan H, Tian J, Tang Y. A shadow imaging bilinear model and three-branch residual network for shadow removal. IEEE Transactions on Neural Networks and Learning Systems. 2023.

[19] Wang L, Zhang L, Yu J, Cheng D, Yao M, Wu H, et al. Relaxed Relational Knowledge Distillation for Inter-Intra Class Learning in Human Activity Recognition. IEEE Sensors Journal. 2024.

[20] Yao M, Zhang L, Cheng D, Qin L, Liu X, Fu Z, et al. Revisiting large-kernel cnn design via structural re-parameterization for sensor-based human activity recognition. IEEE Sensors Journal. 2024;24(8):12863-76.

[21] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. Advances in neural information processing systems. 2017;30.